

DOCUMENT RESUME

ED 191 950

UD 020 841

AUTHOR Gersten, Russell M.: And Others
TITLE Measuring Implementation of a Structured Educational Model in an Urban Setting: Some Preliminary Process-Outcome Findings.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE May 80
GRANT G007507234
NOTE 29p.; Paper presented at the Western Psychological Conference (Honolulu, HI, May, 1980).
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Evaluation Methods; Program Evaluation; *Program Validation; *Site Analysis; *Teaching Models
IDENTIFIERS *Project Follow Through

ABSTRACT

The focus of this report is the Direct Instruction Supervision Code, developed to assess the degree of implementation of the Direct Instruction Follow Through Model. Most of this paper deals with the development, reliability, concurrent validity, and criterion-related validity of the code, which is a classroom observation instrument. The report describes a Follow Through program assessment that employed the code, supervisor ratings, and level-of-use interviews. Findings indicate correlations between degree of implementation and gains in reading. Tables of data are appended. (MK)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

JUL 28 1980

ED191950

Measuring Implementation of a Structured Educational Model
in an Urban Setting: Some preliminary process-outcome findings¹

Russell M. Gersten
Paul Williams
Doug Carnine

University of Oregon

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Russell M. Gersten

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN-
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRE-
SENT OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY

Paper presented at the Western Psychological Conference,
Honolulu, Hawaii, May, 1980

W020841

Footnote

1. This research was supported in part by U.S. Office of Education Grant Number G00-7507234. Without Susan Green and Dr. John Rodriguez of the Department of Education, there could not have been a project of this scope. Ralph Green helped immensely with coordination between the schools and the research staff. Adrienne Allen kept the entire venture together.

Abstract

Degree of implementation of the Direct Instruction model was assessed by classroom observations, supervisor ratings, and Level-of-Use interview. Correlations between degree of implementation and gains in reading were generally in the high-moderate range. (Only one of the five critical reading behaviors-use of corrections-was not highly correlated with academic gains.) Technical issues in process-outcome research are delineated.

Most analyses of the National Follow Through study (Bock, Stebbins, & Proper, 1977; House, Glass, McLean & Walker, 1978) reported unexpectedly large degrees of variability in each model's outcome data. Several authors (House et al, 1978; Kennedy, 1978) have hypothesized that this variability was due in large part to fluctuations in the extent to which a model was implemented at a given site. Yet no attempt was ever made to actually measure fidelity of implementation at the site or classroom level.

In many ways the Direct Instruction Follow Through Model is an ideal candidate for an implementation study. An almost fully articulated system exists (e.g. Becker, 1977; Becker, Engelmann & Carnine, 1979) specifying in detail what teachers should be doing, saying (highly structured formats), how much time should be spent on reading, arithmetic, oral language each day; even how fast lessons should be paced, and at what accuracy rate students should be responding. Thus, operationalization of a behavioral coding system was much more direct and simple than with, say, some of the open classroom models analyzed by the Stallings (1975) group in their Follow Through classroom research. Another unique component of the model is that the role of classroom aides is clearly specified--they actually teach groups of six to ten children curriculum material, oral language, and arithmetic. Thus the same observation system can be used to analyze performance of both teachers and paraprofessionals.

Though several instruments were utilized in the course of the study, the major focus was on the Direct Instruction Supervision Code (Gersten, Allen & Paine, Note 1), a classroom observational instrument. The focus of this paper will be on the development of the code, its reliability (generalizability) and concurrent validity, and initial findings suggestive of criterion-related validity. Because of the relatively small sample size, most of the data analyses are exploratory and tentative in nature.

Nonetheless there seems some evidence that a viable observational code to assess fidelity of treatment can be developed.

Evolution of the Direct Instruction Supervision Code

Perhaps the most crucial step in the development of fidelity of implementation instrument(s) is the specification of which teacher (or staff) performance variables are critical or essential to the full implementation of the model, and which can be adapted or modified without destroying the core concept. Very often, this is done on an a priori basis. This was the case in the early years of the Direct Instruction Model, when, on the basis of their field experience, the model developers offered fairly distinct definition of what model implementation consists of (e.g. Becker, Carnine, Engelmann & Maggs, 1979). While this method has a common sense appeal, there is no empirical support. In fact teachers who are ignoring certain model components--and thus considered low level implementers--may in fact be more effective teachers of children than those teachers who follow every specified step (see Loucks, Newlove & Hall, 1975 for further discussion).

The next step in the specification of the model became a series of empirical studies designed to see whether the 7 specified desired teaching behaviors (see Figure 1) are functionally related to increased student learning. Many of these studies utilized single subject methodology

Insert Figure 1 about here

(Hersen & Barlow, 1976), usually multiple baseline designs, where one to three teachers were studied in depth over a three week period and functional relationships were determined between aspects of teacher performance and student performance variables (i.e. rate of response,

accuracy of response, proportion of students in the group who responded). Three of the implementation variables specified above have been studied individually--the importance of rapid pacing (Carnine, 1976; Carnine & Fink, 1976), correction procedures (Carnine, in press, a; Siegel, Note 2), use of hand signals (Cowart, Carnine & Becker, 1976), systematic verbal praise (Kryzanowski, 1976), and as a package of techniques (Carnine, in press b). Maintaining high student accuracy rate has been a target of research by Brophy & Evertson (1976) and others. The importance of high levels of academic engaged time has been explored by Rosenshine (1977) and Marliave, Fisher, Filby & Dishaw, (Note 3). The only variable in the set which was not examined was the importance of teacher's accurately using all the Distar formats.

The major limitation of the above studies were the relatively brief duration (on the average, two to three weeks) and extremely small sample size (often as small as three students). Furthermore, since in reality these variables act in concert, rather than in isolation, the ecological validity (Snow, 1975; Gage, 1977; Brunswik, 1956) of research would be higher if studies were done looking at actual classroom teachers in a natural setting over the course of the entire school year.

It thus seemed desirable to explore relationships between 7 teacher performance variables and student academic growth in a natural setting. A code was developed, the D.I. Supervision Code (Gersten, Allen & Paine, Note 1), to record rates and frequencies of the following 6 critical behaviors:

1. Accuracy of Formats
2. Use of Hand Signals (to insure unison group response)
3. Use of Corrections
4. Pacing of Lesson

5. Student Accuracy Rate During the Lesson
6. Reinforcement. (For details of the code see Gersten, Note 4, Gersten & Carnine, Note 5).

In addition, supervisors recorded a measure of teacher's time allocation for reading language and math. Detailed rationales for the selection of these variables appear elsewhere. Concurrent validity was assessed by contrasting performance scores with: a) global supervisor rating of fidelity of treatment (on 1 to 4 scale), b) score on the Levels-of-Use instruments (Loucks et al, 1975), a widely used instrument to assess degree of implementation of an education model techniques (Carnine, in press, b). Maintaining high student accuracy rate has been a target of research by Brophy and Evertson (1976) and others.

METHOD

Subjects and Setting

Twenty-one teachers and twenty-one aides were involved in the program. All were regular staff members of either a kindergarten (N = 10) or first grade (N = 11) Follow Through classrooms in one of Follow Through's seven schools in a large urban district. All teachers and paraprofessional aides attended monthly inservice workshops and were periodically observed and offered feedback by on-site supervisors (approximately once a week) and program consultants from the University of Oregon (approximately once every three weeks).

Though the potential sample size was 21 teachers and 21 aides, not every teacher or aide was observed in each phase for a variety of reasons. One of the 7 building principals refused to let the teachers in his school be observed, thus lowering the potential sample size to 17. Some teachers and aides were missed due to illness, or scheduling conflicts. In the

spring, the decision was made to intensively observe a smaller, representative sample of teachers and aides, with each receiving two to four observations within the same week in May, rather than attempt to observe the entire sample of 42 for a mere one-session each. The rationale for this decision is discussed below.

Procedure

Teachers and aides were observed once in November (N = 14 teachers, 15 aides), and once in either late December or January (N = 13 teachers, 14 aides). Observations in May covered fewer teachers but were more intensive to increase the validity of the observational data (N = 13 teachers, 10 aides). Budget limitations precluded more extensive observations.

Teachers/aides were observed in their classrooms while teaching the DISTAR lessons in reading or language to small groups of four to ten students. The observers were familiar to both the teaching staff and the students.

Each teacher (or aide) was observed for 12 minutes teaching a small group DISTAR lesson in either reading (for the teachers) or oral language (for the aides). Observers were trained DISTAR consultants. Each teaching trial was recorded in one frame with 3 components--teacher presentation--student response--teacher response (see Figure 2). Any errors teachers/aides made in lesson presentation (e.g. inaccurate format) or response (failure to correct an error) were noted. The accuracy of student unison responses was also recorded. If even one student failed to respond when a unison response was required, the box was marked NR (No Response). At the end of 12 minutes, the teacher's pace (number of learning trials per minute) and accuracy of the other 4 variables (in percentages) could be calculated. Interobserver reliability was calculated for every fifth observation during the fall and winter sessions.

Dependent measures for the process-outcome correlations were May post-test scores on relevant subtests of the Comprehensive Test of Basic Skills (CTBS) (1974) and gain scores on the WRAT (Wide Range Achievement Test) (1965) using the class as the unit of analysis. Unfortunately no pretest data was available for the CTBS. The use of residual scores (using the WRAT as a predictor) was considered, but the low pre-post correlations (in the .2 to .4 range) indicated no meaningful regression line could be developed. A somewhat similar phenomenon was observed by Cooley and Leinhardt (1980), who also used the class as the unit of analysis.

RESULTS

Reliability of Direct Instruction Supervision Code

Reliability

Interobserver reliabilities on the 5 variables ranged from $r = .80$ to .92 for a random sample of 7 teachers observed by both staff members in the fall and winter. This was deemed adequate.

Temporal stability is a more complex matter. As Shavelson and Dempsey (1976) observe, in order to obtain reasonably stable estimates of individual differences between teachers, each teacher must be observed for many instances during a short period of time. In the fall and winter, each participating teacher/aide was observed only once. Therefore, the fall and winter observational data are not reliable enough to afford valid statements about individual differences. Yet by a simple extension of sampling theory, the estimates of mean teacher (or aide) performance should be unbiased and sensitive to improvements in performance of the samples.

In the spring, a smaller sample of teachers and aides were observed either two or four times within the same week. An exploratory mini-"generalizability" analysis was done to look at variance explained by

occasion of observation error versus the stable variance component corresponding to individual differences in performance. Test-retest correlation were computed between the first and second observation; the Spearman-Brown corrected (for $N = 2$ occasion) correlations are reported in Table 2. The values for Total Score ($r_{tt} = .64$) and Format Accuracy ($r_{tt} = .73$) are quite high for an observational. Pacing and Signals are reasonable at .52 and .46 respectively. Our hope is that, in the 1979-80 study, when each teacher is observed six times within a week, these values will approach the .8 to .9 range. The low value for Student Accuracy may partly be due to a design flaw in the prior observation plan. There was no system to which ability groups (hi, middle, low) the teachers were working with at the time of observation. This has been rectified in future observation schedules. The low coefficient for corrections is puzzling.

For the small sample of five teachers observed four times during one week, coefficient alphas were calculated, treating each observation occasion as a test item. The r_{tt} s were high for all variables, ranging from .6 to .82, although the sample size was so small that no serious generalizations can be made.

In summary, there is evidence of moderate temporal stability for these measures. It is clear that each adult must be observed from four to six (or even eight) times in a one-week time period before stable estimates of his/her performance can be obtained. There must be a systematic sampling of child ability groups. The cost of such an undertaking is appreciable . . . but quite likely worthwhile.

Descriptive Statistics

Table 1 reports descriptive statistics for all three observational periods for the entire sample. Table 2 segregates teachers from aides, offering an overview in trends in skill acquisition.

By November, after two months of inservice training, two skills are mastered by virtually all teachers and aides--use of DISTAR formats, and use of signals. For Format Accuracy, the teacher's mean is 91.6 percent (median = 99.58). For paraprofessionals, the mean is 99.5 percent and the median is 99.9 percent. This is well above established mastery criteria, and little progress is shown during the remainder of the year. The same basic pattern emerges for use of signals, except that some aides take longer to master this skill; $\bar{X} = 95.44$ and $\bar{X} = 83.6$ in November for teachers and aides respectively. Use of formats and signals seem to be routine, mechanical skills that virtually all teachers and most aides acquire within eight weeks.

On the other hand, use of correction procedures, pacing and maintaining student accuracy at a high level appear to be much more subtle, harder to acquire techniques. Acceptable levels determined by DISTAR experts are: 85 percent accuracy for corrections, mean student accuracy of 80 percent, and a pace of 9 trials per minute. November means are well below this level: $\bar{X} = 42.3$ percent for corrections, 60.6 percent for student accuracy, and pace of 6.96 trials per minute for the paraprofessionals. November means are similar for teachers. One can chart steady growth through January until May for both groups. By May the mean teacher pace is 10.1 (mastery level), and corrections is close to mastery ($\bar{X} = 83.3$, $md = 33.5$). Student accuracy level during DISTAR lessons is also close to the acceptable level; $\bar{X} = 78.1$ percent. The paraprofessionals' average scores are slightly lower on all 3 variables--8.36 for pacing, 75.8 for corrections, and 71.2 for student accuracy--indicating a somewhat lower acquisition rate. These three skills definitely appear to require more sophistication, and more supervised training to master.

A striking change is apparent in the use of Allocated Academic Time in the area of reading. In February only 47 percent of the teachers were spending the required time(at least 30 minutes per day) on oral reading (SD = 51). In May, the figure rose to 82 percent (SD = 35). Allocation of time was a key objective in inservice supervision during the winter and spring.

Concurrent Validity

Performance scores on each variable during the Winter Observation Period were correlated with Global Implementation Ratings. Each teacher was independently ranked into quartiles representing the degree of model implementation by the two project consultants. The mean ranking was considered his/her global rating (aides were not ranked). Correlations between these scores on each observational score is reported in Table 1. The median correlation was .45. Correlations were low for only two variables--Reinforcement (which was subsequently dropped from the code) and Signals (a variable which may soon be dropped from the code, see below). For the other 4 variables, significant correlations were found, especially for Pacing (.70), Student Accuracy (.68) and Corrections (.51).

Construct Validity: Preliminary Findings

The increased reliability of the Spring Observational Data allowed for some preliminary attempts of construct validity studies. Two procedures were utilized, (a) contrasted groups (Anastasi, 1976) in which the scores of high and low implementing teachers were explored, and (b) an analysis of the relationship between implementation indicators derived by DISC scores with those derived by administration of the Levels-of-Use interview (Loucks, Newlove & Hall, 1975).

Contrasted Groups

The three highest-ranked and four lowest-ranked teachers in the same group were contrasted on Spring Observational scores, a somewhat tentative approach to construct validity, but given the limited sample size (and limited budget), the only one feasible (see Table 2). There were suggestive differences, tending to favor the high-ranked teachers, on Pacing (md = 11.9 vs 9.5), student accuracy (81.5 vs. 70.8), and especially corrections (md = 83.5 vs. 53.3). Formats showed no difference (95.1 vs. 96.85), probably because all teachers had mastered this most basic skill. Signals showed a difference seeming to favor the low ranked teachers (md = 92.1 for the high group vs. 98.9 for the low group). Interestingly enough, this supports some views (e.g., Meyer, 1979) that use of hand signals is a nonessential component of the Direct Instructional Model; the same effect can be achieved by vocal inflection. Applying appropriate correction procedures and maintaining a high rate of student accuracy during the lessons are certainly more sophisticated.

Relationship between Levels-of-Use and DISC

The Levels of Use (Loucks, Newlove & Hall, 1975) interview was administered to all 21 teachers and 21 instructional aides in February by an independent subcontractor, Emrick and Associates. This interview is a relatively widely used instrument for the assessment of educational innovations; it has both clinical and research utility. We wished to explore the relative pros and cons of using a general instrument like Levels-of-Use versus model-specific instruments such as DISC.

Levels-of-Use protocols were scored by a trained rater. The results indicated that all 21 instructional aides and 16 of the 21 teachers received a score of 3, indicating mechanical use of the innovation. Mechanical users

are teachers still concerned with problems in the mechanics, or technical details of the innovation, i.e., in this case how to follow formats, reinforce correctly, manage the classroom, correct student errors, etc. Five teachers were considered in advanced stages; three received scores of 4A, indicating routine use of the innovation, and two were considered 4B, indicating refined use. Routine users are those who have mastered all the essentials of the innovation, and appear to the interviewer to do so in natural, competent ways. Refined users have mastered the essence of the innovation, and are now adapting to meet the needs of individual students on specific academic projects.

Because of the restricted range of scores, either Pearson or Spearman correlations seemed an inappropriate means of relating Levels-of-Use to DISC scores. Instead, a "profile" for each of the 5 advanced users was created, examining their spring observational scores on each performance variable, and their supervisor ratings. An unusual pattern emerged; of the five, two had the highest performance scores in the district, and one had one of the very lowest. One of these teachers was never observed because she followed the district's bilingual program of teaching Spanish only to non-English speaking children during the first year. (She used no DISTAR, thus not implementing the program at all.)

After reviewing the supervisor reports and DISC scores further, the following became clear. Two teachers rated as 4A by Levels-of-Use, were in fact "Routine," highly competent practitioners of the innovation, performing well above mastery level on every DISC variable and receiving the highest supervisor rating. The third "Routine" user scored below mastery on every DISC variable; she rarely used DISTAR, though she was intellectually knowledgeable about the principles. Thus, the "Refined"

Users included the one teacher who never followed the model, though she was considering so doing (and in fact has implemented it in her classroom in the 1979-80 school year). The other "Refined" user reached mastery level on all but one performance variable, and received a reasonably high supervisor rating.

Several problems with Levels-of-Use should become apparent. The primary one is the lack of variability in the measure; 16 teachers and 21 aides received the same score, though they represented a wide range of levels of implementation. There is also a validity issue; Levels-of-Use interviews seem to detect those with exceptional patterns, but it is possible that knowledgeable, highly verbal users may be able to "talk a good line" even when their day-to-day implementation is uneven, or, in some case, virtually non-existent.

Criterion-Related Validity: Some Preliminary Findings

One of the ultimate goals of the study was to explore the relationship between the six specified teaching behaviors and student academic growth. As previously mentioned, several technical problems were encountered. To briefly review them:

- a) When the instructional group, or class is used as the unit of analysis, sample size is greatly reduced.
- b) to insure adequate temporal stability for the DISC variables, each teacher/aide ideally would be observed four to six times within a one week time frame. This is quite expensive. In the spring a compromise was reached, in which approximately 50 percent of the instructional staff would be observed for two to four occasions within the same week. Unfortunately, this reduced the sample size.
- c) For process outcome studies, each grade level must be analyzed separately. Thus sample size is even further reduced to:

	Teacher	Aides
Kindergarten	7	5
Grade 1	6	5

Before correlations were calculated, the descriptive statistics were calculated for each of these (outrageously small) subsamples. The means and standard deviations for each subsample were compared to those for the entire sample in Table 4, to see whether the samples selected were idiosyncratic, particularly if the range of teacher performance was restricted. Table 4 shows the descriptive statistics for the 4 samples. It is clear that in two of the four cases (Kindergarten teachers and Kindergarten aides) the range is so severely restricted that it is unlikely that any meaningful correlations could be examined.

For the sample of teachers, one finds very restricted range for the 7 Kindergarten teachers in three of the five variables - Use of Corrections, Signals, and Formats. These teachers seem an above group - with higher mean scores than the first grade group, and mean performance at the mastery level for all variables except Accuracy. One must take the restricted range into account in considering this sample. The fact that the observer was not allowed into one school, which contained several low implementing teachers may help explain this phenomenon. The teachers observed in grade 1 seem to represent a broader range of performance.

The samples of instructional aides actually observed are even smaller - Language is assessed by the Test of Language Development (Newcomer & Hammill, 1977), the subtests of the ACLC (a receptive language test) and the DISTAR Continuous Progress Test and Lesson Day in Program. The ACLC subtests are (a) vocabulary, (b) concepts with 2 components (labelled ACLC 2), (c) concepts with 3 components, and (d) concepts with 4 components (ACLC 4). All subtests

seemed relevant for Kindergarten, but only ACLC 4 made sense for first graders. The other subtests were too easy, with all mean scores close to 100 percent. The content relevance of both the TOLD and ACLC are considered low to moderate. (See Gersten, Note 4), when compared to the DISTAR criterion-referenced tests, 5 at each grade level. The Kindergarten aides are consistently higher performers than the first grade aides, and one may again accept some attenuation of correlation for this sample. The range of scores for first grade aides is also limited for the variables Pacing and Formats.

As previously mentioned, mean class post-test (May) scores are used as the dependent measure except for the WRAT Reading, when both post-test and gain scores are used. Reading is assessed by relevant subtests of the Comprehensive Test of Basic Skills (Letter Sounds in K, Letter Sounds, Passage Comprehension, and Word Recognition in Grade 1) as well as a DISTAR criterion-referenced test (called Continuous Progress Test or CPT's) and the Wide Range Achievement Test.

The exploratory correlation matrices are presented in Tables 5 through 8 . With such small Ns, they are at best tentative. Both Pearson and Spearman correlations are calculated (to control for the potential over-weighting of others with Pearson correlations). The first grade teacher matrix shows some consistently positive relationships, for all variables but Pacing and Format Accuracy. The total DISC score is positively related to all variables. The picture for kindergarten teachers is inconsistent and chaotic - restricted range may be a problem.

Several positive relationships are found for the Kindergarten Aides, but the Grade 1 pattern is chaotic. It appears that a contrasted groups analysis may be a better procedure to pursue for secondary analyses. Nonetheless the consistent positive correlations for 2 of the 4 samples

offer a sign that given a larger N, important relationships between teacher performance and student gain may emerge.

Table 4

Descriptive Statistics of Samples
for Process Outcome Study

		Teachers		Aides	
		Mean ($\bar{X}$)	SD	$\bar{X}$	SD
Pace	All	9.8	1.4	8.36	3.25
	K only (N=7)	9.7	1.6	9.0	1.5 ^a
	First only (N=6)	9.9	1.2	7.5	1.6 ^a
Corrections					
	All	83.3	14.3	75.8	30.2
	K only	89.9	9.4 ^a	88.5	9.0 ^a
	First only	75.5	15.8	63.2	28.8
Signals					
	All	94.6	11.6	93.5	16.0
	K only	95.7	8.4 ^a	99.6	.39 ^a
	First only	93.3	15.3	87.4	17.2
Formats					
	All	96.0	7.2	98.6	4.5
	K only	97.9	3.8 ^a	99.8	.40 ^a
	First only	93.8	9.7	97.4	2.9 ^a
Student Accuracy					
	All	78.1	10.8	71.2	11.0
	K only	78.8	9.7	76.2	10.3
	First only	77.3	12.9	66.2	19.0

^aSD much smaller for this grade than total sample

Table 5 |

Preliminary Process Outcome Correlations, Grade I Teachers (N=6)

<u>DISC</u>	Letter Sounds	CTBS Post-test		Word Recognition II	WRAT Gain (N=5)	DISTAR Reading Reading - Mean Continuous Progress Test Score
		Compre- hension	Word Recognition I			
Pacing	.17	-.01	.09	.21	-.21	.51 ^a
Signals	.79*	.52 ^a	.61 ^a	.69 ^a	.48	.92*
Formats	-.37	-.35	-.43	-.68	-.31	-.42
Mean Student Accuracy	.69*	.64 ^a	.68 ^a	.78*	.56	.94*
Corrections	.60 ^a	.65 ^a	.61 ^a	.70 ^a	.63 ^a	.50 ^a
TOTAL DISC	.78*	.62	.66 ^a	.74*	.41	.50 ^a
Allocated Time	.78*	.69*	.74*	.41 ^a	.43 ^a	.43 ^a
Supervisor Rating	.65*	.50*	.55*	.44 ^a	.13	.27

* $p < .05$ ^a Suggestive of positive significant relationship, $.15 < p < .05$

Table 6
FIRST GRADE AIDE CORRELATIONS (N = 5)

	TOLD Oral Vocab.	TOLD Sentence Imitation	ACLC- Subtest 4	Criterion Ref. test in Lang.
PACING	-.65	-.50	-.55	.87*
SIGNALS	-.19	-.43	.30	.02
FORMAT ACCURACY	.57	-.20	.45	-.22
STUDENT ACCURACY	-.54	-.27	-.52	.67
CORRECTIONS	.13	-.05	-.05	.09
TOTAL DISC SCORE	-.31	-.31	-.23	.50
				* = p < .05 + = p < .005
				** = p < .01 ++ = p < .001

Table 7

KINDERGARTEN TEACHER CORRELATIONS
(N = 7)

	CTBS Letter Sound	Reading CPT	WRAT Reading Post-test	WRAT GAIN
PACING	-.47(-.46)	.09(-.09)	.39(-.03)	-.80(-1.0) ⁺⁺
FORMAT ACCURACY	-.65(-.34)	.17(-.02)	-.43(-.34)	.12(-.26)
STUDENT ACCURACY	-.32(-.26)	.46(-.03)	.54(.41)	-.24(-.80)
SIGNALS	-.81(-.85)	.59(.67)	-.13(-.80)	.32(-.11)
CORRECTIONS	-.43(-.29)	.19(.01)	-.62(-.46)	-.02(-.40)
TOTAL DISC SCORE	-.89(-.77)	.51(.23)	.05(-.21)	-.41(-.80)

++ = $p < .001$

Table 8
KINDERGARTEN AIDE CORRELATIONS
(N = 5)

	TOLD Oral Vocab.	TOLD Sentence Imitation	ACL- Vocab.	ACL Subtest 2	ACL # 3	ACL # 4	Criterion Ref. test in Lang.	Day In Program	CT&S List for info.
PACING	.79	.78	.63	.98**	.91*	.85	.28	-.01	.94
FORMAT ACCURACY	.66	.71	.47	.92*	.87	.74	.04	.43	.95
STUDENT ACCURACY	.70	.71	.87	.48	.54	.74	.22	.92*	.04
SIGNALS	.60	.05	.38	.57	.27	.42	.45	.32	.85
CORRECTIONS	.13	.45	-.07	.56	.63	.29	-.48	-.01	.91
MEANTOTM	.75	.88	.63	.96*	.78*	.86	.11	.50	.85

* = $p < .05$
** = $p < .01$

+ = $p < .005$
++ = $p < .001$

References

- Anastasi, A. Psychological testing (4th ed.). New York: MacMillan Publishing Co., Inc., 1976.
- Becker, W.C. Teaching reading and language to the disadvantaged -- What we have learned from field research. Harvard Educational Review, 1977, 47, 518-543.
- Becker, W.C., Engelmann, S.E., Carnine, D.W., Maggs, A. Direct Instruction technology -- making it happen. Advances in Child Behavior Analysis and Therapy, 1979
- Becker, W.C. Engelmann, S., & Carnine, D. The Direct Instruction Model. In R. Rhine (Ed.) Encouraging change in America's schools: A decade of experimentation. New York: Academic Press, 1979.
- Bock, G., Stebbins, L.B., & Proper, E.C. Education as experimentation: A planned variation model. Vol. IV-A Effects of Follow Through models. U.S. Office of Education, 1977.
- Brophy, J. & Evertson, C. Learning from teaching: A developmental perspective. Boston: Allyn and Bacon, 1976.
- Brunswik, E. Perception and the representative design of psychological experiments. Berkeley, Calif.: University of California Press, 1956.
- Carnine, D. W. Effects of two teacher presentation rates on off-task behavior, answering correctly and participation. Journal of Applied Behavior Analysis, 1976, 9, 199-206.
- Carnine, D.W. Correcting word identification errors of beginning readers. Education and Treatment of Children. In press a.
- Carnine, D.W. High and low implementation of direct instruction teaching techniques. Education and Treatment of Children. In press b.

- Carnine, D.W. & Fink, W.T. Increasing rate of presentation and use of signals in elementary classroom teachers. Journal of Applied Behavior Analysis, 1978, 11, 35-46.
- Cooley, W.W. & Leinhardt, G. The instructional dimensions study. Education Evaluation and Policy Analysis, 1980, 2, 7-25.
- Cowart, J.B., Carnine, D.W., & Becker, W.C. The effects of signals on child behaviors during DISTAR instruction. University of Oregon Follow Through Project Technical Report 76-1, Appendix B, 1976, 55-85.
- Gage, N. The scientific basis of the art of teaching. NY: Teachers College Press, 1977.
- Herson, M. & Barlow, D.H. Single case experimental designs: Strategies for studying behavior change. New York: Pergamon Press, 1976.
- House, E.R., Glass, G.V., McLean, L.D., & Walker, D.F. No simple answer: Critique of the "follow through" evaluation. Harvard Educational Review, 1978, 28(2), 128-160.
- Kennedy, M. Findings from the Follow Through Planned Variation Study. Educational Researcher, 1978, 7 (6), 3-11.
- Krysanowski, J.A. Praise effects of on-task behavior during small group instruction. University of Oregon Project Follow Through Technical Report 76-1, Appendix B, 1976, 29-37.
- Loucks, S.F., Newlove, B.W., and Hall, G.E. Measuring Levels of Use of the Innovation: A Manual for Trainers, Interviewers, and Raters. University of Texas at Austin, R&D Center for Teacher Education, Austin, Texas, 1975.
- Meyer, L. Teaching basic skills to low income children: One community's approach. Unpublished manuscript. Eugene, OR: University of Oregon, 1979.

- Rosenshine, B.V. Academic engaged time, content covered and direct instruction. Paper presented to the A.R.E.A. Conference, New York, 1977.
- Shavelson, R. & Dempsey, N. Generalizability of measures of teaching behavior. Review of Educational Research, 46(4), 553-611.
- Snow, R.E. Representative and quasi-representative designs for research on teaching. Review of Educational Research, 44, 265-291.
- Stallings, J.A. Implementation and child effects of teaching practices in Follow Through classrooms. Monograph of the Society for Research in Child Development, 1975, 40, (7-8, Serial No. 163).

Reference Notes

1. Gersten, R., Allen, A., & Paine, S. Direct Instruction supervision code. Unpublished manuscript. Eugene, OR: University of Oregon, 1980.
2. Siegel, M. An experimental investigation of teacher behavior and student achievement in the DISTAR instructional system. Unpublished doctoral dissertation, University of Illinois, 1973.
3. Marliave, R., Fisher, C., Filby, N. & Dishaw, M. The development of instrumentation for a field study of teaching. Technical Report I-5. San Francisco: Far West Laboratory, 1977.
4. Gersten, R. The San Diego implementation study: Interim report. Technical Report 79-1: University of Oregon, Eugene, OR, 1979.
5. Gersten, R., & Carnine, D. Measuring implementation of the Direct Instruction model in an urban school district: An observational approach. Paper presented at A.E.R.A., Boston, Mass, April, 1980.